DOCUMENT RESUME

ED 310 468 CS 506 805

AUTHOR Halay, Kathryn; Roberts, Charles V.
TITLE The "Watson-Barker Listening Test" for High School Students.
PUB DATE Mar 89
NOTE 17p.; Paper presented at the Annual Meeting of the International Listening Association (10th, Atlanta, GA, March 1-4, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Auditory Stimuli; *Aural Learning; High Schools; *High School Students; *Listening; Listening Comprehension; *Listening Comprehension Tests; Listening Skills; Sex Differences
IDENTIFIERS Test Adaptations; *Watson Barker Listening Test

ABSTRACT

The high school version of the Watson-Barker Listening Test was developed in response to the need for a listening test appropriate for high school students. The test was comprised of conversations that would normally occur in either the high school setting or in the home and was developed in two different versions. The test consists of five sections each of which has ten questions, based on two or more stimuli. The sample included 397 high school students from around the country (218 females and 179 males) aged 13-18, who were asked to complete one or both forms of the test. Males scored consistently lower than females on all parts of each form of the test, a difference in keeping with previously reported gender differences. The standard deviations on all parts, and the total scores, were lower than those reported for the adult video version of the Watson-Barker Listening Test. Although there was great variability among the testing sites, tremendous differences among the subjects, differences in motivation in subjects, and less than rigorous application of standard empirical control during much of the testing, the alternative forms of the test still were found to be significantly correlated, with each form capable of predicting a significant amount of the variation in the alternative form. (Four tables of data are included.) (MG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED310468

The Watson-Barker Listening Test For High School Students

by

Kathryn Halay

&

Dr. Charles V. Roberts

McNeese State University

Lake Charles, Louisiana 70609-0420

Paper Presented at the

Tenth Annual

International Listening Association Convention

Atlanta, Georgia - March, 1989

Running Head: High School Listening Test

CS 506805

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Kathryn Halay, Charles V. Roberts
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Watson-Barker Listening Test For High School Students

by Kathryn Halay, Dr. Charles V. Roberts

Even a cursory look at the average K-12 and/or university curriculum of the past discloses that the least attended communication skill was listening. At the same time, the evidence illustrates overwhelmingly that listening was clearly and consistently found to be the most used communication skill (Rubin & Roberts, 1987). Studies have found that 52 percent of teaching emphasizes reading, while only 8 percent was devoted to listening (Benoit & Lee, 1988). However, adults spend 42.1 percent of their verbal communication time listening, while they spend 15 percent reading (Wolvin & Coakley, 1985). This inverse relationship between use and training creates a dilemma that has caused several educators to wonder whether "... our educational system has been built upside down" (Benoit & Lee, 1988, p. 229). This topsy-turvy state of educational affairs may have been allowed to continue partly because people have not been made aware of how good or bad they listen. Research shows that although people spend a large amount of time listening, they generally do not listen well, yet report that they are superior listeners (Benoit & Lee, 1988; see also Wolvin & Coakley, 1985).

This bias against listening training seems to be changing. One of the positive trends that can be noted in our educational system today is the "turning rightside up" of skills awareness and education both in academic circles and business communities (Roberts, 1988). Business people have begun to realize that ineffective listening causes a decrease in productivity. Because of this, many employers have begun listening training programs for

their executives, office people, and shop workers (Wolvin & Coakley, 1985). The fact that listening can be taught and learned effectively is becoming increasingly accepted and is resulting is substantial changes in academic curricula (Benoit & Lee, 1988). Further evidence of the increased interest being shown concerning listening is the growth of the International Listening Association. Founded in 1979, the ILA promotes the exchange of listening materials and research findings among professionals, and has grown to where it now has members in 41 states, the District of Columbia, and a dozen foreign countries.

One of the possible reasons for such a turn around in thinking is the relatively recent creation of several listening measurement devices. Earlier listening measures have been criticized for a variety of reasons (Roberts, 1988). The newer scales were created to both sensitize individuals to their listening weaknesses and as research devices intended to help develop methods for increasing listening effectiveness. Before effective teaching methods of listening skills can be developed reliable measures of skill levels must be devised. The process that awaits the listening researcher and educator can be illustrated by drawing an analogy to the medical community. In medicine, before a treatment can be devised for an ailment, the disease must first be identified and its various symptoms and parts formally delimited from secondary problems and processes that may be related. Then, after authorities have agreed on what constitutes the disease, the next task is to develop a method for measuring the existence and/or extent of the problem. Finally, researchers may attempt to develop inoculations against the disease and/or remedies for the symptoms of the disease.

The same types of problems confront the listening expert as she attempts to develop "inoculations and antidotes" for the "societal disease" of poor listening. The term "listening", though used often, and apparently consistently commonsensically defined by those everyday users, is not seen as consistently by theorists (Rubin & Roberts, 1987).

The conceptual variations have to do mostly with which sub-processes should be included in the definition of listening. Various scholars have included some of the following processes under the listening label: hearing, perceiving, attending, comprehending, retaining, recalling, and responding appropriately (Rubin & Roberts, 1987). The mix is such that it seems that one could ask five researchers to define and delimit listening and they would respond with six different and distinct definitions of listening. At this point the only agreement among theorists is that listening is a multidimensional concept (Hauser & Hughes, 1987). A resolution of the definitional problem would help shape theories and guide research and serve to allow our attention to focus more directly on how to better listening skills.

While "it would seem more prudent to first discover what it is that we should be studying before deciding on how we should measure it (Roberts, 1988, p. 3)," such an agreement of theorists is not mandatory before progress can be made, and the next step of test development for the individual researcher. Currently, though no single listening test has yet gained universal acceptance, and many definitions of listening continue to vie for the allegiance developed listening tests that purport to reliably and validly measure the listening skills of an individual, as defined by the chosen conceptual definition (Roberts, 1988) Each of these few definitional groupings is engaged in impressive research programs to develop these tests further.

Two of the most popular listening test today are the Watson-Barker Listening Test and the Kentucky Listening Test. Both tests were developed in the early 1980's as standardized listening tests that were oriented primarily toward adults and mature college level audiences (Roberts, 1988). Both have been used for pedagogical research purposes as well as to sensitize people as to the need for listening training. Though both uses are essential, the latter use logically would temporally precede the former. As was noted earlier, people

typically are unmotivated to work on listening skills and businesses are unlikely to budget funds for training sessions without first knowing that they have a listening weakness.

Both tests are purported to be valid. based on large part because of their "face validity." Such claims are dependent upon the conceptualization of the underlying construct. If agreement does not exist among researchers as to what constitutes the domain of a listening test, then there can be little hope that these same theorists would agree to the validity of the instrument. Roberts (1988) provided another method for "Severing the Gordian Know" of validity claims and supported the validity of the Watson-Barker Listening Test, while at the same time criticizing other validity problems. Though the Watson-Barker Listening Test does include paralinguistic cues and does measure an individual's ability to decode these signals, it, of necessity, neglects other nonverbal cues present in most listening situations. In many listening settings, the aural verbal message is also accompanied by various nonvocalized, nonverbal cues that travel through other channels (Roberts, 1988).

Regardless of the definitional boundary set, it is readily apparent that none of the test creators would be able to sustain the validity of their instrument for other audiences. The appropriateness of using a test developed for adults on children, or even children of high school age would be questioned, especially since there is a noted situational/content bias in both of the most popular tests that orients them to situations that are unfamiliar to many children. This should concern us. Some maintain that "secondary students become progressively more ineffective listeners as they progress through the secondary school" (Benoit & Lee, 1988). If students at this level could be taught to listen effectively, their futures in college, and therefore in the working world, would be better off. Effective measurement is key to effective pedagogical attacks on listening deficiencies. As of 1988, there were no extant listening tests designed to measure the listening abilities of high school students.

Realizing the importance of a reliable and valid measure of listening intended to measure the abilities of high school students, the creators of the Watson-Barker Listening Test undertook the task of adapting their instrument for the high school audience in 1988. A pilot video version of a high school test were developed. The video format was selected for several reasons. "Videotape has been found to be more involving, and learning has been positively related to the two-channel mode more than to a one-channel mode" (Rubin & Roberts, 1987). Roberts (1988) has suggested that video listening tests have a greater face validity than do audio listening tests, though this opinion is not shared by all theorists in the field (Bostrom & Waldhart, 1988). Finally, an audio version of the final test is planned. Creating an audio tape from a video stimuli is relatively straightforward and would allow a greater measure of reliability than would attempting to create a video version that would approximate an audio tape.

The Development of the High School Version of the Watson-Barker Listening Test

The adult version of the Watson-Barker Listening Test was developed in 1982 to measure recognition of stimulus material via audiotape (Watson & Barker, 1984). The test was intended for adults (18 years old or older). A video version of the audio test was developed in 1987 to measure receiver aural and visual decoding activities (Roberts, 1987). Both the audio and video versions were oriented to adult audiences in terms of knowledge references and language choice.

The increasing call for a listening measurement instrument appropriate for the high school audience prompted the creators of the Watson-Barker Listening Test to undertake a high school version of the test (Watson & Barker, 1988). Preliminary work began in around 1986 to adapt the adult version to suit a high school audience. The adult version was

comprised of conversations that might occur in common adult situations, usually having to do with either the work or college environment (Watson & Barker, 1988). In the high school version conversations were developed that would normally occur in either the high school setting or in the home. Word choice was restricted to the vocabulary level of the high school freshman, and references were restricted to those topics and situations it was reasoned would be familiar to the high school student.

Preliminary scripts of two alternative forms of the listening tests were examined by high school students at various locations around the country. Their reactions to the various potential test stimuli were used to guide further refinements of the scripts. This process was repeated a second time with the revised script, and still more refinements were incorporated into the script. As the planning progressed, consideration of criticisms of the adult version led to certain restrictions concerning character selection.

There are five sections that comprise the test. Part 1 measures how well the individual can evaluate message content. Part 2 measures the degree to which the individual understands meaning in conversations. Part 3 measures how well test takers understand and remember information presented to them in lectures. Part 4 tests the ability of individuals to evaluate emotional meaning in messages. The final section, Part 5, tests how well the subject can follow instructions and directions. These parts are intended to mimic the adult version of the Watson-Barker Listening Test. Each part has ten questions, based on two or more stimuli. Scores are arrived at by computing correct answers and multiplying the sum of these correct responses by two. Thus the potential range of scores on each of the five section is 20. In an effort to increase the reliability between the two alternative forms, the same number of women, men, young men and women, and equal representation of the same ethnic groups were planned for both versions in each of the five sections. Each version of the test takes approximately 40+ minutes to complete.

The adult version of the listening test has a diverse array of dialects represented in the various stimuli. This attempt to represent a cross section of the dialects found in the United States has been criticized by some as creating more adjustment problems than the typical listener would encounter in everyday life. Eastern audiences, in particular, had reacted negatively to what they perceived as an abundance of southern accents. It was decided that rather than attempt to use diverse dialects, actors/actresses would be chosen who could speak using a standard American dialect.

Since the test was to be videotaped, further planning concerning settings and characters was undertaken. Settings were chosen for their familiarity for high school students. These were designed to look like high school classrooms and home scenes. The people who were to generate the stimuli were of the appropriate age to fit the parts of young and "old" speakers. Each participant was given his/her script in advance of the taping, was asked to dress in regular clothing, that would not either disclose the region in which it was produced, or draw undo attention to the wearer.

During the taping some slight alterations were made as suggested by the high school aged individuals being taped. Several versions of each stimulus were taped. Immediately after the taping the various scenes were reviewed. Several technical problems were discovered in a few of the scenes, and these were re-taped. The variations of each scene were inspected and the one that best approximated the ideal was chosen. In several cases the answers to specific questions were altered to conform to the stimuli.

An initial face validity check was run in the summer of 1988. Twenty students enrolled in a speech class a southern university were shown the completed tape and were asked to verify the appropriateness of each answer, given the chosen stimuli. Special emphasis was

given to Section Four which assesses the ability of individuals to read the nonverbal cues present in a message. Several problems with consistency were discovered and final adjustments were made to the answer key. Edited versions of two alternate forms of the high school listening test were prepared.

Testing the Pilot Version

The pilot versions of the Watson-Barker High School Listening Test were tested at locations around the United States. The sample was drawn from a variety of socio-economic and geographic areas. In all, 397 high school students, aged 13 to 18 were asked to complete one or both forms of the test. There were 218 female students and 179 male students of varying grade point averages and grade levels in the sample. A total of 259 students from the larger sample were given both forms of the test. After the tests were administered, they were graded as indicated in the instructions and the resultant scores analyzed.

First, descriptive statistics were generated for the total sample and, because gender differences had been noted in previous tests of video-taped listening tests (Roberts, 1987), descriptive statistics also were generated for the two sub-samples. The following tables (I and II) illustrate the total sample's mean scores and standard deviations for each part and the total test, and mean scores for each gender sub-sample for each part of each form and the total score of each form.

Table I

Form A (N=320)

Part	Mean Score (all)	Mean Score (137 males)	Mean Score (183 females)	St.Dev.(all)
1	16.25	16.04	16.40	2.75
2	17.56	16.89	18.06	2.56
3	16.91	16.38	17.30	2.93
4	17.20	16.93	17.40	2.03
5	14.28	13.96	14.58	3.71
TOTAL	82.19	80.12	83.74	8.59

Table II

Form B (N=341)

Part	Mean Score (all)	Mean Score (148 males)	Mean Score (193 females)	St. Dev.(all)
1	17.47	17.31	17.60	3.00
2	16.11	15.72	16.41	2.94
3	15.58	14.97	16.04	3.41
4	15.97	15.46	16.35	2.92
5	15.62	14.82	16.24	3.44
TOTAL	80.62	78.28	82.42	11.22

As can be seen, there were differences between both Form A and Form B and between genders. Males scored consistently lower than females on all parts of each form of the test. This consistent difference is in keeping with previously reported gender differences. There is some variation among the scores of the corresponding parts of Form A and Form B. Note that the mean scores on Parts 1 and 5 of Form A are lower than the corresponding mean scores of Form B, while the mean scores for Parts 2, 3, and 4 were higher for Form A than for Form B. The total score of A is higher than that of Form B.

The standard deviations for on all parts and the total scores were lower than those reported for the adult video version of the Watson- Barker Listening Test. While this difference may be due to the larger sample size (at least 320 versus 98 - see Roberts, 1987), the ranges reported below indicate that it also could be attributable to a ceiling effect.

Combining the weighted scores of the two forms resulted in the normative data in Table III.

Table III

Forms A and B Combined

Part	Mean Score (all)	Mean Score (males)	Mean Score (females)
1	16.88	16.70	17.02
2	16.81	16.28	17.21
3	16.22	15.65	16.65
4	15.57	16.17	16.86
5	14.97	14.41	15.43
TOTAL	80.62	79.16	83.06

Additional analyses were undertaken using only those subjects who had completed both forms of the test.. A series of Pearson Product Moment correlation tests were conducted to ascertain the alternate form reliability of the two versions of the test. For this latter group, an attempt was made to alternate the order of presentation of the two forms so that approximately half of the subjects would take Form A before taking Form B and vice versa. Table IV illustrates how the various sections and the total scores were correlated and the range of correct responses for each part and for the total scores of both forms.

As the analyses indicate, the two forms of the test are correlated, and the correlations of the various sub-parts are higher than those reported for the adult version of the Watson-Barker

Listening Test (Watson & Barker, 1988). The weakest correlation for any part was for Part 4. This section taps the ability of subjects to read the non-verbal aural and visual cues present in the message. Though there is some variability, an a priori $p < .1$ level of confidence was set for the pilot version and this section met that criterion level.

Table IV

Forms A and B (N=259)

Part	Form A Mean (Range)	Form B Mean (Range)	r	p <
1	16.52 (8-20)	17.71 (0-20)	.26	.0001
2	17.75 (8-20)	16.15 (8-20)	.17	.006
3	17.03 (6-20)	15.93 (4-20)	.22	.0008
4	17.32 (4-20)	16.23 (4-20)	.11	.09
5	14.65 (0-20)	15.85 (0-20)	.38	.0000*
TOTAL	83.26 (46-98)	81.88 (38-98)	.53	.0000*

*Denotes a probability <.00009

One other analysis was undertaken to see if the scores on the various parts of Form A would significantly predict the total score of Form B and vice versa. The scores of the 259 subjects who took both forms of the test were analyzed using the multiple regression technique. A multiple correlation of .5425 ($r^2 = .2944$, $p < .0009$) was found for the five parts of Form A regressed on the total score for Form B and a multiple correlation of .5695 ($r^2 = .3243$, $p < .0009$) was found for the five parts of Form B regressed on the total score for Form A. In both cases the largest Beta weights were for sections 1 and 5. Approximately one-third of the variation in the test is predictable by scores on the alternative Form.

Discussion

A video test specifically designed to tap the listening skills of the high school subject has been needed for many years. That no such test has been created before is easily understandable to anyone who has tried to create such a test. That the two forms of this test yield correlated scores is heartening. That the correlation of the total scores and the correlations of the parts are weak, is understandable given the variability found among the test conditions. It is encouraging that the multiple regression analysis reveals that the variations in the scores can be significant predicted by the part scores on the alternative version.

The array of subject ability, using other predictors such as GPA and socio-economic background could best be described as bi-modal, with most of the subjects being found at the two ends of the continuum. This same distribution could account for the possible ceiling effect noted earlier. That the correlation holds up with affluent, upper level socio-economic subjects mixed in with inner city, lower socio-economic class subjects is remarkable and argues forcefully for the utility and generalizability of the instrument.

The conditions under which the pilot testing was undertaken were not without distractions and were not uniform. The authors of this paper did administer some of the tests, but other tests were administered by helpful high school teachers around the country. Some of the tests were administered by teachers interested in listening, while others administered the test primarily because they were asked to do so and had no deep interest in listening. Perhaps as a result, student attention and motivation was reported at different levels at different testing sites. In unsolicited comments that were sent in with the completed tests it was revealed that "some students were non-attentive" and that others "had to be whipped into line." Others were kinder noting that "some students ...were not as diligent about

taking the test as they should have been." One very helpful administrator told of how one teacher reported that his students "fooled around quite a bit and considered the test 'silly' and 'too stilted.'" If all were as unmotivated, then this explanation of the variability would not hold as much merit since all would be equally unmotivated. Fortunately a number of the subject took the task seriously and seemed to listen as well as they possibly could. Other variation was noted as well. The time between testing the two forms varied. Some subjects took the tests with only one day intervening, while others waited a whole week between events. Some had their test scores from the first test reported to them before taking the second test, while others did not. Some students reported that they found taking the second test easier because they had "learned" how to take the test. This comment came from students who had taken both Form A and Form B first. Finally, some students thought that they had taken the test previously, noting this comment on their answer sheet. That they were mistaken gives evidence of the similarity of the instruments, and the lack of retention of those same students.

Of course, no researcher should be praised for not limiting the variability among test conditions, though a case can be made for purposefully investigating the robustness of a measure. The argument for the robustness of the test is not one that should be dismissed simply because it was not the intent of the researcher. The fact remains that though there was great variability among the testing sites, tremendous differences among the subjects, differences in motivation in subjects, and less than rigorous application of standard empirical control during much of the testing, the alternative forms still were found to be significantly correlated, with each form capable of predicting a significant amount of the variation in the alternative form.

Certainly changes can be made to future versions of the Watson-Barker Listening Test, Form A and Form B. Questions could be made more difficult to decrease the likelihood of

a ceiling effect. Instructions could be made more forceful so that teachers would know to increase the motivation of students, hopefully equalizing it at a level that would insure the tapping of the optimal listening skill level of the student. But this instrument is a start. Even in it present form it can serve to sensitize students as to their relative level of listening skill. It can be used in its current form to give us some direction in how effectively we are teaching our high school students to listen more effectively.

BIBLIOGRAPHY

Benoit, S. & L. Lee. "Listening; It can be taught!" Journal of Education for Business, 63 (February 1988): 229-32.

Bostrom, R. & E. Waldhart. The Kentucky Comprehensive Listening Test. Lexington: Kentucky Listening Research Center, 1980.

Bostrom, R. & E. Waldhart. "Memory models and the measurement of listening." Communication Education, 37 (January 1988): 1-13.

Hauser, M. & A. Hughes. "A factor analytic study of four listening tests." Journal of the International Listening Association , 1 (Spring 1987): 129-47.

Roberts, C. "The validation of listening tests: Cutting of the Gordian Knot." Journal of the International Listening Association, 2 (1988): pp. 1-19.

Rubin, R. & C. Roberts. "A comparative examination and analysis of three listening tests." Communication Education, 36:2 (April 1987): pp. 142-53.

Steil,L., L. Barker, & K. Watson. Effective Listening: Key to Your Success, Reading, Massachusetts: Addison-Wesley Publishing Company, 1983.

Watson, K. & L. Barker. "Listening assessment: The Watson-Barker Listening Test." Journal of the International Listening Association, 2 (1988): 20-32.

Watson, K. & L. Barker. Watson-Barker Listening Test. New Orleans: SPECTRA, Inc., 1984

Wolvin, A. & C. Coakley. Listening. Dubuque, Iowa: Wm. C. Brown Company Publishers, 1985.